Historic, Archive Document

Do not assume content reflects current scientific knowledge, policies, or practices.

UNITED STATES DEPARTMENT OF AGRICULTURE
Agricultural Marketing Service

INTERPRETATION OF VARIABILITY IN THE DATA USED IN PREPARING THE COTTON GRADE AND STAPLE REPORTS

(Revised)

An Office Report

Prepared by
F. H. Harper
In July 1934, while a member of the Staff
of the
Bureau of Agricultural Economics

Washington, D. C.
October 1939

Interpretation of Variability in the Sample Data Used in Preparing the Cotton Grade and Staple Reports

by

F. H. Harper

\- - - - - -

During the past decade a particularly keen interest has developed in the possibilities of improving sampling methods and in evaluating the adequacy and reliability of sample data selected to represent stratified universes. Numerous contributions have been made in this branch of statistical technique, and there has been marked advancement in the appraisal and application of different sampling procedures. The preparation of this paper purports to indicate some of the advantages of refinement in the analysis of variability in paired data, to demonstrate application that has been made of results obtained from analyses of variability in the allocation of samples used as a basis of reports on the grade and staple length of the American cotton crop, and to indicate how the variability in the sample data used in preparing the reports has been evaluated and interpreted.

The analysis of problems involving quality of cotton produced in the United States is facilitated by the use of reports on the grades and staple lengths of cotton ginned. The usefulness of these reports, issued by the United States Department of Agriculture, and

now available for the entire domestic crops of 1928 to 1933, inclusive, will depend somewhat upon the extent to which agricultural workers and others using them acquaint themselves with the data and the procedure followed in assembling them.

In 1927, Congress enacted legislation (Act of March 3, 1927, Public No. 740 - 69th Congress--44 Stat. 1372-1374) authorizing and directing the Secretary of Agriculture to collect and publish annually "statistics or estimates concerning the grades and staple lengths of stocks of cotton on hand on the 1st of August of each year in warehouses and other establishments of every character in the continental United States," and also to publish "not less than three such estimates" with respect to each crop. It is further provided that "In any such statistics or estimates published, the cotton which on the date for which such statistics are published may be recognized as tenderable on contracts of sale of cotton for future delivery under the United States Cotton Futures Act of August 11, 1916, as amended, shall be stated separately from that which may be untenderable under said Act as amended."

The Division of Cotton Marketing, Bureau of Agricultural Economics, inaugurated in 1928, in compliance with the provisions of the Act, the preparation of reports on the grade, staple length, and tenderability of the entire domestic crop, funds for the purpose becoming available for the fiscal year beginning July 1, 1928. 1/ Preliminary work of

1/ The first report on the 1928 crop was issued on September 28, 1928, and had reference to cotton ginned in the United States prior to September 1. The first report prepared in compliance with the Act on the "carry-over" of American cotton was issued on September 21, 1928, and referred to stocks of cotton on hand in the United States at the beginning of the cotton marketing year on August 1.

preparing reports on the grade, staple length, and tenderability of ginnings had already been undertaken in 1927 by the Division of Cotton Marketing, Bureau of Agricultural Economics, under the general authority contained in the Agricultural Appropriation Act of January 18, 1927, and with funds made available for the fiscal year beginning July 1, 1927. 2/

Partly because the available funds were inadequate for a survey of the entire Cotton Belt, and partly because the work was not inaugurated early enough in 1927 to permit the establishment of an organization sufficient for that purpose, the inquiry pertaining to the 1927 crop, or the 1927-28 season, was of restricted scope, covering only the State of Georgia, 20 counties in northern Texas, 3/ and 7 counties in the contiguous part of southwestern Oklahoma. 4/

Prior to the inauguration in 1927 of this preliminary work by the Division of Cotton Marketing, some earlier attempts had been made to furnish information on staple length of cotton produced and ginned in the United States. 5/ The Bureau of Crop Estimates 6/ of the United

2/ The first report of the 1927-28 season was issued on October 4, 1927, and had reference to cotton ginned in the State of Georgia prior to September 1. Funds for carrying on the work were provided from those originally alloted to the Division of Crop and Livestock Estimates.

3/ Including Baylor, Childress, Collingsworth, Cottle, Dickens, Foard, Hall, Hardeman, King, Knox, Motley, Wichita, Wilbarger, Crosby, Donley, Floyd, Hale, Hockley, Lamb, and Lubbock counties.

4/ Including Comanche, Cotton, Greer, Harmon, Jackson, Kiowa, and Tillman counties.

5/ Results of early measurements of cotton fiber lengths were published by the Census Office of the Department of the Interior in Volume 5 of the 1880 Census, pages 14 to 35, inclusive.

6/ The Division of Statistics, established May 28, 1863, became the Bureau of Statistics in 1903, and this Bureau became the Bureau of Crop Estimates on July 1, 1914. The Bureau of Crop Estimates was combined with the Bureau of Markets on July 1, 1921, to form the Bureau of Markets and Crop Estimates. On July 1, 1922, the Office of Farm Management and Farm Economics was combined with the latter to form the Bureau of Agricultural Economics, at which time the Division of Crop and Livestock Estimates was established.

States Department of Agriculture, organized in 1914, prepared estimates on the staple length of cotton ginned from the 1914 crop. 7/ The inquiry concerning the length of staple specified long staple as being 1-3/16 inches and upward in length. 8/ In preparing estimates concerning the staple length of the 1915 crop, long staple cotton was considered to be that having a length of 1-1/8 inches and more. The preparation of staple-length estimates 9/ was discontinued in 1926 by the Division of Crop and Livestock Estimates, which from the time of its establishment in 1922 had continued the staple-length estimates inaugurated by the Bureau of Crop Estimates. The last of these prepared reports for the United States had reference to the 1925 crop. 10/

7/ Farmers' Bulletin 651, February 6, 1915, pp. 12-13, and Cotton Production and Distribution, Bureau of the Census, Bulletin 1831, page 24.

8/ Monthly Crop Report, June, 1916, page 50, and Farmers' Bulletin 651, February 6, 1915, page 12.

9/ Information relative to these estimates is readily accessible as follows: 1914 crop, Farmers' Bulletin 651; 1915 crop, Monthly Crop Report, June, 1916, page 51, and Monthly Crop Report, June, 1917, pp. 52-53; 1916 crop, Monthly Crop Report, June, 1917, pp. 52-53, and United States Department of Agriculture Bulletin 733, pages 7 and 8; 1917 crop, United States Department of Agriculture Bulletin 733, pages 7 and 8; 1918 crop, Monthly Crop Reporter, June, 1920, page 52; 1919 crop, Monthly Crop Reporter, June, 1920, page 52, and Monthly Crop Reporter, April, 1921, page 45; 1920 crop, Monthly Crop Reporter, April, 1921, page 45. Published records, if any, of estimates pertaining to the crops of 1921 to 1925, inclusive, are not available.

10/ Reports were prepared by L. L. Janes on the 1926 and 1927 crops of Louisiana, but they were not published. North Carolina and Arkansas are two states for which the furnishing of information on the quality of the cotton crop was attempted prior to the inauguration of reports by the Division of Cotton Marketing. (See United States Department of Agriculture Bulletin 476 for information relative to classification of samples from the North Carolina crops of 1914 and 1915, and Arkansas Extension Circular 92 for results of classification of samples from the Arkansas crops of 1915, 1916, 1917, and 1918.)

In order for the Division of Cotton Marketing to carry out the provisions of the Act of March 3, 1927, it seemed appropriate to procure samples of cotton from different localities in each of the cotton-producing states. Prior to the inauguration of this work there was no comprehensive information indicating differences in grade and staple length of cotton produced and ginned in these states, and for this reason the procuring of samples could not at first be planned on the basis of stratification according to known differences in the cotton. The importance of proper stratification soon became apparent to the field representatives, however, when it was realized that the number of samples on which to base the grade, staple-length, and tenderability reports would necessarily be limited by funds and facilities available for complying with the statute. 11/

In order to procure samples on which to base the grade, staple-length, and tenderability reports, the Division of Cotton Marketing makes arrangements each year with certain ginners in different parts of the cotton-producing states whereby they agree to furnish a sample of about four ounces from each bale ginned at their gins, the sample to be drawn from the gin press box or from the bale after it has been pressed and

11/ The total number of samples of upland cotton classed and used in preparing reports on specified crops is as follows: 1928, 1,359,254; 1929, 1,007,952; 1930, 1,139,763; 1931, 1,032,401; 1932, 841,579; 1933, 739,544.

The total number of samples of American-Egyptian cotton classed and used in preparing reports is as follows: 1928, 3,980; 1929, 4,334; 1930, 4,638; 1931, 2,459, 1932, 1,894; 1933, 3,773.

tied. To date, practically all of the samples classed for purposes of quality statistics have been press-box samples, but a few ginners have cut them from the bales.

When arrangements were made with ginners for the procurement of samples during the 1928-29 season, no comprehensive information was available that would reliably indicate variations in the grade and staple-length of cotton produced from one state to another or in different localities within the individual states. In view of this fact, gins were selected that season to represent states without reference to these variations and without any particular reference to geographic heterogeneity. An attempt was made, however, to insure the procurement of an aggregate of samples in each state sufficiently large to provide a safe margin of adequacy and reliability with respect to representative variability, it being realized at the time that the aggregate of samples likely to be procured might be greater in some instances than absolutely necessary to insure reliable reports on the grade and staple-length of total ginnings.

After the data pertaining to the classification of samples had become available for the entire ginning season of 1928-29, as a preliminary step to assembling them in final form, the producing states were stratified into districts, based in a general way on prevailing soil types. This division into districts (See U. S. Dept. Agri. Statistical Bulletin 40), boundaries of which were adjusted to conform to county lines, not only permitted

the assembling of the grade and staple data on the basis of convenient subdivisions and the comparison of variability among them, but it obviated the possible necessity of determining a series of weights in order to properly emphasize the relative importance of each grade and staple length ginned in different parts of producing states, especially in those instances in which the aggregates of samples procured in different parts of producing states may not have been uniformly proportionate to total ginnings therein.

The data pertaining to the grade and staple length of ginnings of the 1928-29 season showed considerable variability, as was expected, from one state to another and even within the individual states. It was logically inferred, therefore, in view of the size of sample procured, that there were greater degrees of heterogeneity in some states and localities than in others. Because of this varying heterogeneity, the indications (to field representatives) obviously were that the proportionate distribution of the aggregate of samples in relation to ginnings should be different from one state to another, from one division of a state to another, and, in many instances, even from one part of a division to another. This, of course, is obvious. It is mentioned, however, in order to place the proper emphasis upon the fundamental problem involved in the sampling of grade and staple universes made up of a large number of heterogeneous stratifications.

Distinction has been carefully made in sampling cotton ginnings between "uniform variability" and "representative variability," and the analyses upon which this report is based are predicated on the assumption that, at least so far as the mathematics is concerned, the statistical heterogeneity is an indicator of needed representation and sample requirement. It is realized that there is no rigid formula or equation by which the number of samples or gins required to insure representativeness in the proportions of the different grades and staple lengths can be accurately determined. Procedure has been devised and adapted, however, by which a logical allocation of an aggregate of samples is made in accordance with the extent to which stratifications and larger universes are homogeneous. Statistical technique, involving the analysis of variability in the sample data from one season to another, has been found useful in appraising the adequacy and reliability of sample data procured as representations of stratified grade and staple-length universes.

With final assembling of the data pertaining to the 1929 crop

at the end of the 1929-30 ginning season, there were available the figures representing number of samples of each grade and staple length that had been received from cooperating ginners during two consecutive seasons. It became possible, therefore, to arrange paired series of observations for those gins from which samples had been received both seasons and to analyze the variability both between and within these series.

For purposes of obtaining numerical measurements of variability in the sample data, and in order to provide for an evaluation of the relative importance of variability in different producing areas, data pertaining to the staple length of cotton ginned at cooperating gins have been analyzed in their entirety and by arranging them in paired groups. That part of the procedure pertaining to the derivation of squared variability (increments of the standard deviation squared) is so flexible that it permits the separation of variability estimated as having been contributed from different detected sources. 12/ The value of statistical method in studying variability in staple-length distribution of cotton samples received from the cooperating gins is occasioned by the existence and effects of numerous influencing

12/ See "The Analysis of Variance Method of Measuring Differences between Staple-Length Designations of Press- Box and Cut Samples of Cotton," October, 1933, by F. H. Harper and W. B. Lanham. (Mimeographed.) See also other reports mentioned in footnote 25.

heterogeneous factors. 13/ Among these is geographic location of cotton-producing areas, some of which are characterized by a wide diversity of soil, cultural practices, and weather conditions.

The adoption of a stratified sample, beginning with the 1929-30 season, by dividing the states into districts was intended to increase the homogeneity of individual universes and thus facilitate the proper proportionate sampling thereof. 14/ This constituted the first attempt made since the inauguration of the project to gain mastery in the sampling procedure over the relation between geographic location and grade and staple length of cotton produced and ginned.

The need for stratification within districts of producing states has also been recognized, and it has led to an increasingly careful study each year of grade and staple-length variability and of geographic heterogeneity. This study has appropriately been divided into two phases. One of these consists of analyses of differences in cotton varieties

13/ There is analytical technique admirably adaptable also to the measurement of differences in paired and replicate classifications of identical samples of cotton and of different samples from the same bale which makes possible an evaluation of the relative importance of variability contributed from different detected sources. It is adaptable also in many instances to the analysis and interpretation of price differences.

14/ Lanham, W. B., and Harper, F. H. Jour. of Farm Economics, Volume XVI, Number 2, April, 1934, pages 329-333. For the 1928-29 season, the Cotton Belt was stratified according to state boundaries. For the 1929-30 season, however, the individual states were stratified into districts, the boundaries of which were adjusted to county lines, the primary purpose being to provide universes that would possibly be less heterogeneous than entire states. (See U. S. Dept. Agri. Statistical Bulletin 40.)

planted, soil heterogeneity, ginning machinery, and other factors, the possible effects of which on grade and staple-length variability do not readily lend themselves to precise mathematical measurement. Consideration of these factors and their possible effects has been found useful in allocating the aggregate of samples which it has been possible to procure and class with the funds and facilities available for these purposes.

The other phase of the study consists of analyzing the statistical variability in proportionate staple-length distribution of ginnings from year to year within individual states and districts. 15/ The results of these analyses furnish indications of the difficulty that might be expected in the sampling of universes by providing for the appraisal of the degree of statistical heterogeneity and homogeneity of staple-length distributions. They serve also another useful purpose in the sampling procedure by providing for an estimate of the number of homogeneous, or approximately homogeneous, statistical stratifications.

In making the final analyses of variability in staple-length distribution and the determination of homogeneous, or approximately homogeneous, statistical stratifications, paired data have been used throughout, but data pertaining to all cooperating ginning establishments have also been analyzed in appraising the importance of variability in relation to the sampling procedure. To avoid the use of large numbers

15/ No similar analyses have been made of grade variability.

and, which is more important, to correct for differences in volume of ginnings at identical gins during successive seasons, most of the analyses have been based on the percentages representing the proportions of each staple length ginned rather than on the actual number of bales. It then becomes possible to make a logical comparison and analysis of differences in distribution of ginnings at the same gins.

The following equations indicate some of the basic considerations underlying the analyses of sample data used in preparing the grade and staple-length reports.

1. $\sum[(d_x) + (d_y)]^2 = \sum(d_x)^2 + \sum(d_y)^2 + 2\sum(d_x)(d_y)$

2. $\frac{\sum[(d_x) + (d_y)]^2}{n} = \frac{\sum(d_x)^2}{n} + \frac{\sum(d_y)^2}{n} + \frac{2\sum(d_x)(d_y)}{n}$

3. $\sum[(d_x) + (d_y) + (d_u)]^2 = \sum(d_x)^2 + \sum(d_y)^2 + \sum(d_u)^2 +$
$2\sum(d_x)(d_y) + 2\sum(d_x)(d_u) + 2\sum(d_y)(d_u)$

4. $\frac{\sum[(d_x) + (d_y) + (d_u)]^2}{n} = \frac{\sum(d_y)^2}{n} + \frac{\sum(d_y)^2}{n} + \frac{\sum(d_u)^2}{n} +$
$\frac{2\sum(d_x)(d_y)}{n} + \frac{2\sum(d_x)(d_u)}{n} + \frac{2\sum(d_y)(d_u)}{n}$

If observations in two series are from the same universe, then, on the average, the standard deviation squared of one is expected to equal the standard deviation squared of the other, and the product of the two standard deviations will be equivalent, on the average, to the square of the individual standard deviations. It is apparent, of course, that the squares of standard deviations calculated for series

representing different periods of time need not necessarily be identical in magnitude in order for the indications to be that the different gins or series of gins are sampling the same, or very similar, stratifications, especially if seasonal changes affect the variability.

Whenever $r = 0$, $\sigma^2_{x+y} = 2\sigma^2$, as the following illustration will indicate. The data in columns 2 and 3 of table 1 are used in making this illustrative presentation.

Table 1.- Illustrative data in which the variance of the column of summations equals two standard deviations squared

1	2	3	4	5
Duplicate observation number	x	y	x + y	xy
1	40	10	50	400
2	40	20	60	800
3	40	30	70	1200
4	50	10	60	500
5	50	20	70	1000
6	50	30	80	1500
Total	270	120	390	5400
Mean	45.0	20.0	65.0	900
Mean of squares	2050.00	466.67	4316.67	---
Square of mean	2025.00	400.00	4225.00	---
Standard deviation squared	25.00	66.67	91.67	---

$\sigma x = 5.00$

$\sigma y = 8.1651$

$$\text{Product-moment} = \frac{\Sigma xy}{n} - M_x M_y = 0$$

$$r = 0, \text{ and } \sigma^2_{x+y} = 2\sigma^2 \text{ (i.e., } \sigma^2 x + \sigma^2 y)$$

The coefficient of correlation is the quotient obtained by dividing the correlated variability 16/ by the geometric mean 17/ of the

16/ Correlated variability may be conveniently referred to as "covariance."

17/ The geometric mean is the nth root of a product. When there are two numbers, the geometric mean is the square root of their product; when there are three numbers, the geometric mean is the cube root of the product; etc.

variances. In actual calculation, the product-moment becomes the correlated item, and the product of the two standard deviations constitutes the equivalent of the geometric mean of the variances. An important consideration in the derivation and interpretation of a coefficient of correlation is that there should be a sufficiently large number of observations in the series to overcome the tendency toward unity. It is to be remembered also that errors of observation do not cancel out in obtaining "r", so that the calculated coefficient will not be the same as it would be if these errors were not present.

Separation of the correlated and uncorrelated parts of variability is readily accomplished after the squares of the standard deviations have been obtained. The following equations, in which the symbols "c" and "u" are adopted to designate the correlated and uncorrelated parts of variability, respectively, will illustrate the procedure as applied to table 1. In presenting the results, fractional parts of variability are referred to as standard deviation squared instead of increments of standard deviation squared. Analysts may use the latter term if it is felt that the results referred to as standard deviation squared do not convey sufficient implication in this respect.

$$4\,\sigma^2_c + 2\,\sigma^2_u = 91.67 \text{ (or } \sigma^2_{x+y}\text{)}$$

$$2\,\sigma^2_c + \sigma^2_u = 45.835 \text{ (or } \sigma^2_{x+y} \text{ divided by 2)}$$

$$\sigma^2_c + \sigma^2_u = 45.835 \text{ (or one-half of } \sigma^2_x + \sigma^2_y\text{)}$$

$\sigma^2_c = 0$ (or $45.835 - 45.835$)

$\sigma^2_u = 45.835$ (or $45.835 - 0$)

These calculations are the equivalent of the following:

$$(1)\ \sigma^2_c = \frac{\sigma^2_{x+y}}{2} - \frac{\sigma^2_x + \sigma^2_y}{2} = 45.835 - 45.835 = 0$$

$$(2)\ \sigma^2_u = \frac{\sigma^2_x + \sigma^2_y}{2} - \sigma^2_c = 45.835 - 0 = 45.835.$$

The value 91.67 includes, as the equations indicate and as the calculations have illustrated, four parts of σ^2_c, which is a zero quantity in this instance, and two parts of σ^2_u. One-half of this value, therefore, or 45.835, includes two parts of σ^2_c and one of σ^2_u. The average of the sum of 25.00 and 66.67, the squares of the standard deviations of the two individual series, includes one σ^2_c and one σ^2_u, and σ^2_{x+y} is equal to σ^2_x plus σ^2_y plus $2\ \sigma^2_c$.

Any difference remaining after one-half the summation of the squares of standard deviations of the two individual series (columns 2 and 3) is subtracted from one-half the standard deviation squared of the x + y series (column 4) is attributable to the fact that the latter value contains one σ^2_c more than does the former value.

The difference between these two measures represents, therefore, the magnitude of that extra σ^2_c. Its derivation is one of the calculations incident to the solving of two normal equations. In this particular instance the two values are identical, since there is no correlated item of variability, and the difference between them must necessarily be zero. It is apparent, then, that the total variability is accounted for by the uncorrelated item.

In order to evaluate the effect of season, or the relation that season bears to the observed variability, it is desirable to break down the uncorrelated item into its component parts. Seasonal changes and their effects on staple-length variability are not controlled, of course, under field conditions of cotton production, and it is for this reason that there is need of a measure for that part of uncorrelated variability that is in addition to the contribution attributable to season. There is then available for use in further interpreting the extent of homogeneity, as well as heterogeneity and resulting stratification, some indication of the relation that the uncorrelated variability contributed by differences "between the series" bears to the uncorrelated variability contributed by differences "within the series."

There might be instances in the analysis of certain types of data in which the correlated item will account for the total calculated variability, since if $r = 1$, then $\sigma^2_{x+y} = 4\sigma^2$. The relation that this, together with the fact that if $4 = 0$, $\sigma^2_{x+y} = 2\sigma^2$, bears to analyses by which the correlated and uncorrelated parts of variability are separated is appreciated when the probability is realized of both parts of variability frequently, if not generally, occurring in the same total. It would likely be a very rare exception under actual conditions of sampling, especially in the case of biological populations, if an instance were found in which either the correlated or uncorrelated part of variability accounts for the total variability in observations pertaining to successive seasons. The following example will illustrate perfect correlation, thus showing that when $r = 1$, $\sigma^2_{x+y} = 4\sigma^2$.

Table 2.- Illustrative data in which the variance of the column of summations equals four standard deviations squared

1	2	3	4	5
Duplicate observation number	x	y	x + y	xy
1	12	12	24	144
2	18	18	36	324
3	25	25	50	625
4	30	30	60	900
5	40	40	80	1600
Total	125	125	250	3593
Mean	25	25	50	718.6
Mean of squares	718.6	718.6	2874.4	---
Square of mean	625.0	625.0	2500.0	---
Standard deviation squared	93.6	93.6	374.4	---

$\sigma_x = 9.6747$

$\sigma_y = 9.6747$

$$\text{Product-moment} = \frac{\Sigma xy}{n} - M_x M_y = 93.6$$

$$r = \frac{\text{product-moment}}{\sigma_x \sigma_y} = \frac{\frac{\Sigma xy}{n} - M_x M_y}{\sigma_x \sigma_y} = 1.0,$$

and $\sigma^2_{x+y} = 4\sigma^2$ (i.e., $\sigma^2_x + \sigma^2_y$ times 2)

As in the first illustration, in which the coefficient of correlation is equal to 0, r is calculated by dividing the covariance 18/ by the geometric mean of the variances. The covariance, corresponding to the product-moment, is of the same magnitude in this instance as the geometric mean of the variances, which is the equivalent of the product of the two standard deviations. When this agreement in magnitude occurs, it is obvious that the coefficient of correlation must necessarily be perfect.

Determination of the correlated and uncorrelated parts of variability is accomplished by the procedure and equations already presented in connection with table 1, in which, as before, the symbol "c" is used to designate the correlated item and the symbol "u", the uncorrelated item. We have, therefore, the following:

$$4\,\sigma^2_c + 2\,\sigma^2_u = 374.4 \text{ (or } \sigma^2_{x+y})$$

$$2\,\sigma^2_c + \sigma^2_u = 187.2 \text{ (or } \sigma^2_{x+y} \text{ divided by 2)}$$

$$\sigma^2_c + \sigma^2_u = 93.6 \text{ (or one-hald of } \sigma^2_x + \sigma^2_y)$$

$$\sigma^2_c = 93.6 \text{ (or } 187.2 - 93.6)$$

$$\sigma^2_u = 0 \quad \text{(or } 93.6 - 93.6)$$

These calculations are the equivalent of the following:

$$(1)\quad \sigma^2_c = \frac{\sigma^2_{x+y}}{2} - \frac{\sigma^2_x + \sigma^2_y}{2} = 187.2 - 93.6 = 93.6$$

$$(2)\quad \sigma^2_u = \frac{\sigma^2_x + \sigma^2_y}{2} - \sigma^2_c = 93.6 - 93.6 = 0.$$

18/ The covariance will not be expected to be larger than the geometric mean of the variances, since the coefficient of correlation calculated by the ordinary methods may range only from 0 to 1.

The correlated part of variability is 93.6, and the uncorrelated part is 0. Four times the correlated part of variability plus two times the uncorrelated part equals 374.4, as stated in the first of the preceding equations and σ^2_{x+y}, column 4 of table 2, is equal to σ^2_x plus σ^2_y plus $2\sigma^2_c$.

The two illustrations represented by tables 1 and 2 have indicated the conditions under which total variability would be expected to be accounted for entirely by either the correlated or the uncorrelated item. In actual sampling problems, these conditions, as already indicated, would seldom, if ever, be expected to occur, especially in biological populations pertaining to successive seasons. A comprehension of possibilities becomes more essential, therefore, in order to facilitate the interpretation of the constituent elements of total variability, to which these correlated and uncorrelated parts contribute.

It is by such interpretation of variability that the analyst concerned with the sampling of stratified universes is often able to formulate logical conclusions relative to the homogeneity of the population from which the aggregate of samples was drawn. This, together with some logical appraisal of the extent and importance of heterogeneity, constitutes one of the principal problems in sampling a universe.

The Grade and Staple Statistics Section of the Division of Cotton Marketing has relied completely on sample data in preparing the grade and staple reports on cotton ginned in the United States during the six seasons of 1928-29 to 1933-34. 19/ It has been important,

19/ See United States Department of Agriculture Bulletin 40, "Grade, Staple Length, and Tenderability of Cotton in the United States, 1928-29 to 1931-32"; and periodic grade and staple releases, 1928-29 to 1933-34.

therefore, to make some study of the variability in ginnings within districts and states from year to year and to interpret as far as practicable the causes of differences in proportionate distribution of ginnings at the same gins and in the same localities from one year to another. Variability in staple length of ginnings has been found to be so great in many instances that no attempt has yet been made with the sample now being procured to report on the quality of cotton ginned in each individual county or on the quality of cotton ginned in producing areas smaller than the districts into which the states are subdivided. 20/

In analyzing the variability in distribution among the different staple lengths of ginnings at cooperating gins during consecutive seasons, the mathematical technique has been designed to furnish results that indicate the homogeneity and heterogeneity of the statistical

20/ See footnotes 11 and 14. It is apparent, of course, that by sampling every bale at certain gins, as the cooperating ginners agreed to do, a larger aggregate of samples was probably obtained in some instances than was absolutely necessary to represent a community or locality. So far, however, it has not been considered advisable to sample only a part of the bales ginned at cooperating gins. In the first place, sampling only a part of the bales might result in the samples being drawn at improper intervals, and, in the second place, this plan would result in information being available on the quality of the cotton belonging to some patrons of a gin and not on the quality of the cotton belonging to other patrons. It will be realized, however, that if such a plan were feasible the number of stratifications could be increased without any change in the aggregate of samples. Ultimately it may be considered desirable by the Congress to sample and class all cotton bales ginned at all gins.

The Association of Southern Agricultural Workers in convention at Jackson, Mississippi, February 5, 6, and 7, 1930, adopted a resolution asking for "adequate appropriations to provide a larger statistical sample in all of the states." (Proceedings, 31st annual convention, pp. 5 and 6.)

universes sampled. The following illustration represents one procedure that may be useful in some instances in the analysis and interpretation involved in the isolation of correlated variability 21/ without the determination of the magnitude of individual parts contributing to the uncorrelated variability. Gins represented in table 3 are a part of those that cooperated with the Division of Cotton Marketing during the seasons specified by furnishing samples to be classed and used in preparing the grade, staple-length, and tenderability reports.

21/ In this paper the expression "standard deviation squared" is used for convenience throughout instead of "increment of standard deviation squared" in referring to fractional parts of variability contributing to the total. Analysts may use the latter term if it is considered preferable.

Table 3. - Variance analysis of the percentage distribution of cotton shorter than 7/8 inch ginned at six Louisiana gins during specified successive seasons

Gin designation 1/	Percentage distribution of cotton shorter than 7/8 inch 2/		
	1928-29 (x)	1929-30 (y)	x + y
A	46.1	33.3	79.4
B	34.9	43.9	78.8
C	12.6	15.2	27.8
D	2.4	2.9	5.3
E	37.4	71.8	109.2
F	16.4	20.6	37.0
Total	149.8	187.7	337.5
Mean	24.97	31.28	56.25
Mean of squares	862.38	1475.86	4434.73
Square of mean	623.50	978.44	3164.06
Standard deviation squared	239.08	497.42	1270.67

1/ Representing gins in Lincoln, Bienville, Sabine, DeSoto, Union and Claiborne parishes.

2/ The percentages for the individual seasons represent the proportions that cotton shorter than 7/8 inch was of the total ginned at specified gins.

$$4\sigma^2_c + 2\sigma^2_u = 1270.67 \text{ (or } \sigma^2_{x+y})$$

$$2\sigma^2_c + \sigma^2_u = 635.34 \text{ (or } \sigma^2_{x+y} \text{ divided by 2)}$$

$$\sigma^2_c + \sigma^2_u = 368.25 \text{ (or one-half of } \sigma^2_x + \sigma^2_y)$$

$$\sigma^2_c = 267.09 \text{ (or } 635.34 - 368.25)$$

$$\sigma^2_u \text{ for } 6 = 101.18 \text{ (or } 368.25 - 267.09)$$

$$\sigma^2_u \text{ for } 12 = 101.16 + 9.92 \text{ (or } 101.16 + \sigma^2 \text{ of}$$

[illegible]

In this problem the calculations have been facilitated by carrying decimals to only two places. If decimals were carried one additional place in divisions made subsequent to the calculation of the squares of standard deviations, then one-half of 1270.67, the square of the standard deviation of $x + y$, would be expressed as 635.335 instead of as 634.34, in accordance with the engineer's rule, which operates to prevent all the inaccuracies being in one direction. The calculated correlated item would be 267.085 if decimals were carried this additional place, and the uncorrelated item would be 101.165 instead of 101.16. In this instance, four times the correlated item plus two times the uncorrelated item would be equal to 1270.67, and σ^2_{x+y} would be equal to σ^2_x plus σ^2_y plus $2\ \sigma^2_c$. The carrying of decimals in such instances to more than two places solely for the purpose of obtaining a greater degree of precision in the arithmetic calculations may not be warranted because of certain characteristic variability in the basic data.

It is obvious in this instance, because of the variability, that there is stratification within the series. The indications of stratification are readily observed, of course, independent of the calculations, because of the wide range in magnitude of observations. It is because of this wide range, together with the consequent indicated certainty of stratification, that the x and y distributions in table 3 are analyzed for purposes of illustrating the technique that has been applied in determining the extent of statistical stratification represented by the sample data used in preparing the grade and staple-length reports. The circumstances which permit the detection of

significant stratification, such as is apparent in table 3, without the necessity of measuring the variability arithmatically should help to clarify the analytical concepts of the method used in studying differences in staple-length distribution of cotton ginned in the different states and districts.

The correlated item of 267.09 is the equivalent of the product-moment, which can be readily proved by subtracting the product of the two means, 24.97 and 31.28, from the mean product of paired x and y observations. The product of the two means is 781.06, and the mean of the products of the paired observations is 1048.15. Between these two values there is a difference of 267.09, corresponding to the correlated item. The uncorrelated item for six observations is accounted for by the difference between 267.09 and 368.25. This difference is 101.16. For twelve observations the uncorrelated item is 111.08, the sum of 101.16 and the standard deviation squared of the means of x and y. The magnitude of this squared standard deviation is influenced by any seasonal changes which cause a difference between the two means.

The relative magnitudes of correlated and uncorrelated parts of variability are affected by differences in paired observations and by differences within the individual series. In order to further illustrate this relationship and to show in greater detail the statistical constituency of variability in ginnings at the same gins during consecutive seasons, the values derived from the data in table 3 have been recalculated and the decimals carried a greater number of places. This illustration of differences between the magnitudes of numerical values

will further indicate the application of this method of analysis in studying the relationship between fractional parts of total variability inherent in paired series of percentages representing the proportions that cotton of any specified staple length was of total ginnings at gins represented.

[illegible] [illegible]

[illegible] [illegible]

[illegible]

[illegible]

Table 4. - Variance analysis of the percentage distribution of cotton shorter than 7/8 inch ginned at six Louisiana gins during specified successive seasons 1/

Gin designation	Percentage distribution of cotton shorter than 7/8-inch		
	1928-29 (x)	1929-30 (y)	x + y
A	46.1	33.3	79.4
B	34.9	43.9	78.8
C	12.6	15.2	27.8
D	2.4	2.9	5.3
E	37.4	71.8	109.2
F	16.4	20.6	37.0
Total	149.8	187.7	337.5
Mean	24.96667	31.28333	56.25
Mean of squares	862.57667	1475.85833	4434.72833
Square of mean	623.33461	978.64674	3164.06250
Standard deviation squared	239.24206	497.21159	1270.66583

1/ Basic data taken from table 3. Values are recalculated by carrying the decimals a greater number of places than they were carried in table 3.

$$4\sigma_c^2 + 2\sigma_u^2 = 1270.6658 \text{ (or } \sigma^2_{x+y})$$

$$2\sigma_c^2 + \sigma_u^2 = 635.3329 \text{ (or } \sigma^2_{x+y} \text{ divided by 2)}$$

$$\sigma_c^2 + \sigma_u^2 = 368.2268 \text{ (or one-half of } \sigma_x^2 + \sigma_y^2)$$

$$\sigma_c^2 = 267.1061 \text{ (or } 635.3329 - 368.2268)$$

$$\sigma_u^2 \text{ for 6} = 101.1207 \text{ (or } 368.2268 - 267.1061)$$

$$\sigma_u^2 \text{ for 12} = 101.1207 + 9.9750 \text{ (or } 101.1207 + \sigma^2 \text{ of means)} = 111.0957$$

[illegible]

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

[illegible]

The common mean of the x and y observations is 28.125, the quotient obtained by dividing 12 into 337.5, the total summation. The mean of the squares of deviations of individual x and y observations from the common mean is 378.2018. This mean of squared deviations exceeds 368.2268, which includes one correlated part and one uncorrelated part of variability, by 9.9750, the square of the standard deviation of the two means, 24.96667 and 31.28333.

It may be interesting to observe also that the difference between the average of the squares of all the x and y observations and the product of the means of x and y is greater than 368.2268, or the product-moment plus the uncorrelated part of variability, by an amount equal to twice the magnitude of 9.9750, the square of the standard deviation of the means. The difference between the average of the squares of the two means and their product is also equal to twice the square of this standard deviation. This latter difference, that is, the difference between the average of the squares of the two means and their product, exceeds the square of the standard deviation of the x and y means by an amount equal to the product of the deviations of the two individual means from the common mean. 22/ These deviations of individual means from the common mean are necessarily of the same magnitude.

If observations in the y series were of the same magnitude as the observations in the x series with which they are paired, there would be no uncorrelated variability, as was the case with the distributions of

22/ These relationships, well known to mathematicians, are herein mentioned because a complete comprehension of them might assist in interpreting the results of the analytical technique applied to the study of variability in ginnings at identical gins during consecutive seasons.

[illegible]

observations presented in table 2, in which the correlated item accounts for total variability. In such an instance as that represented by paired observations in table 2 the square of the standard deviation of the series of summations of paired x and y observations is equal to four times the square of the standard deviation of either of the individual series, or exactly twice as great as the summation of the two squared standard deviations calculated for the x and y series.

With the correlated and uncorrelated parts of variability isolated, it is then possible to proceed with the separation of uncorrelated variability into its component parts. The isolation of the different parts of variability helps to make it possible to formulate logical conclusions as to the amount of stratification, which is directly related to the sampling procedure. A part of the stratification may be attributable to changes in varieties planted from year to year. The effect of changes in varieties on variability in staple length of ginnings is reflected in the results of the analyses.

A further analysis of this general type is sometimes desirable in determining the relative importance of the different parts of variability contributed from the different detected sources. This analysis is easily accomplished by arranging the data to facilitate a two-way stratification. It is then possible to calculate readily the measures representing each part of variability without any rearrangement of the observations. The following table and calculations are presented to illustrate the technique and to indicate the interpretation to be

placed upon the results.

It will be observed that the fractional parts of variability contributing to the total are referred to as standard deviation squared, just as they have been in the analysis of data in tables 1, 2, 3, and 4. If considered preferable, which it may well be, these fractional parts of variability can be referred to in each analysis as increments of the standard deviation squared. (See footnote 21.)

Table 5. - Variance analysis of the percentage distribution of 7/8-inch cotton ginned by five Alabama gins during specified successive seasons

Gin designation 1/	Percentage distribution of 7/8-inch cotton 1928-29 (x)	1929-30 (y)	x + y	Mean	Mean of squares	Square of mean	Standard deviation squared
A	76.70	64.00	140.70	70.350	4989.445	4949.123	40.322
B	71.61	60.40	132.01	66.005	4388.076	4356.660	31.416
C	68.74	66.90	135.64	67.820	4600.399	4599.552	.847
D	63.04	57.90	120.94	60.470	3663.226	3656.621	6.605
E	86.82	58.20	145.02	72.510	5462.476	5257.700	204.776
Total	366.91	307.40	674.31	337.155	114558.854	113673.494	885.360
Mean	73.382	61.480	134.862	$\sum \sigma^2$ = 283.966			
Mean of squares	5449.565	3791.884	18255.725	Mean of $\sum \sigma^2$ = 56.7932			
Square of mean	5384.918	3779.790	18187.759	σ^2 of means = 16.9914			
Standard deviation squared	64.647	12.094	67.966	σ^2 of total of 10 = 73.7846 2/			

1/ Representing gins in Geneva, Houston, Dale, Henry, and Barbour counties.

2/ If decimals in all the calculations were carried one or more additional places, the σ^2 of a total of 10 obtained by this procedure would be of the same magnitude as the σ^2 of a total of 10 obtained by summating the mean of the squares of the standard deviations of the x and y columns and the square of the standard deviation of the means of x and y.

$$\sum \sigma^2 = 76.741 \text{ (or } 64.647 + 12.094)$$

$$\text{Mean of } \sum \sigma^2 = 38.3705 \text{ (or } \sum \sigma^2 \text{ divided by 2)}$$

$$\underline{\sigma^2 \text{ of means (73.382 and 61.480)} = 35.4144}$$

$$\sigma^2 \text{ of total of 10} = 73.7849 \text{ 3/ (or } 38.3705 + 35.4144)$$

3/ See footnote 2.

$$4\sigma_c^2 + 2\sigma_u^2 = 67.9660$$

$$2\sigma_c^2 + \sigma_u^2 = 33.9830$$

$$\sigma_c^2 + \sigma_u^2 = 38.3705$$

σ_c^2 = - 4.3875 (gin)

σ_u^2 = 42.7580 (season + error for 5 observations)

σ^2 of means = 35.4144

Total = 78.1724 (season + error for 10 observations)

$$9\sigma_c^2 + 3\sigma_u^2 = 885.3600$$

$$3\sigma_c^2 + \sigma_u^2 = 295.1200$$

$$\sigma_c^2 + \sigma_u^2 = 56.7932$$

$2\sigma_c^2$ = 238.3268

σ_c^2 = 119.1634 (season)

σ_u^2 = -62.3702 (gin + error for 2 seasons)

σ^2 of means = 16.9914

Total = - 45.3788 (gin + error for 10 observations)

- 45.3788 (gin + error for 10 observations)

- 4.3875 (gin)

- 40.9913 (error for 10 observations) 23/

78.1724 (season + error for 10 observations)

119.1634 (season)

- 40.9910 (error for 10 observations) 24/

23/ If decimals are carried a sufficient number of places, identical values for error can be obtained by this calculation and by the determination of the difference between the measures of variability for season and for season and error combined. A check on each of these series of calculations is thus provided.

By multiplying each derived part of variability by the number of observations and then dividing by corresponding degrees of freedom, the following is obtained:

Gin	= - 4.3875 x 10,	or	- 43.8750,	÷ 4	= - 10.9688
Season	= 119.1634 x 10,	or	1191.6340,	÷ 1	= 1191.6340
Error	= -40.9913 x 10,	or	-409.9130,	÷ 9	= - 45.5459
Total	= 73.7846 x 10,	or	737.8460		

The value 737.8460 is the equivalent of the measure of total squared variability that may be obtained by squaring the individual x and y observations, summating, and then subtracting the product of the sum of the two series and their common mean. In studies involving the analysis of differences in paired and replicate classifications of cotton samples, variability has been analyzed by the Analysis of Variance method and separated into component parts preparatory to evaluating the degree of significance of differences between variances and preparatory to evaluating the relative importance of "bias" and of that part of "spread" which is in addition to "bias."

With component parts of variability attributable to differences in classing thus obtained, the error is readily separated from each part and proportionate contributions readily determined. 25/ This procedure for analyzing differences in classing is the only one yet known to have be suggested that can be used both in measuring relative magnitudes of parts of variability contributed to the total from different detected sources and in interpreting the degree of significance of differences between magnitudes of contributions from different sources. 26/

25/ See the following:

a. "Variance Analysis of Variability in Paired and Replicate Series of Staple-length Observations on Cotton Samples," by F. H. Harper, W. B. Lanham, And O. T. Weaver (Journal of Farm Economics - July, 1934, Vol. XVI, No. 3, pages 529-530.)

b. "Measurement of Average Differences between Paired Observations on Staple Length of Cotton Samples," by O. T. Weaver, W. B. Lanham, and F. H. Harper (Journal of Farm Economics - July, 1934, Vol. XVI, No. 3, pages 534-535).

c. "The Analysis of Variance Method of Measuring Differences between Staple-length Designations of Press-Box and Cut Samples of Cotton," by F. H. Harper and W. B. Lanham (Mimeographed report issued by the Department in October, 1933).

d. Numerous office reports prepared in the Grade and Staple Statistics Section of the Division of Cotton Marketing on classing differences. Copies of these reports are on file in the library of the Division of Cotton Marketing.

26/ This method of procedure is the best one known for emphasizing the difference between "bias" and "spread" in classing.

According to the calculations following table 5, $4\sigma^2_c + 2\sigma^2_u$ are equal to 67.9660, the square of the standard deviation of the column of $x + y$ values. The calculated correlated item for gin is - 4.3875, whereas the calculated uncorrelated item for error plus season is 42.7580. Four times the correlated item is -17.5500 and two times the uncorrelated item is 85.5160. The total of these two products is 67.9660, which verifies the statement that this quantity represents $4\sigma^2_c + 2\sigma^2_u$, the correlated item being, as observed, a negative quantity.

The results of the second part of the calculations by which the two parts of variability are separated show that $9\sigma^2_c + 3\sigma^2_u$ equals 885.3600, the square of the standard deviation of the two x and y summations, 366.91 and 307.40. The calculated correlated item for season is 119.1634, and the calculated uncorrelated item for gin and error is -62.3702. Nine times the correlated item is 1072.4706. Three times the uncorrelated item is -187.1106. The algebraic sum of the two products is 885.3600, representing $9\sigma^2_c + 3\sigma^2_u$.

After the number of stratifications has been determined, it is possible to logically allocate a given aggregate of samples among the producing states according to the ratio that the figure representing volume or probable volume of ginnings by the gin or gins proposed for each statistical stratification bears to the figure representing the total summation of ginnings by all gins proposed to constitute the

sample. For statements pertaining to determination of stratifications and to the allocation of samples see the following office reports, copies of which are on file in both the Grade and Staple Statistics Section and in the library of the Division of Cotton Marketing.

a. "Analysis of Variability in Staple Length of Cotton Ginned During the Seasons of 1928-29, 1929-30, and 1930-31 by Certain Gins and Size of Sample Calculated for 1931-32," September, 1931, pages 184-193.

b. "Analysis of Variability in Staple Length of Cotton Ginned During the Seasons of 1930-31 and 1931-32 by Certain Gins and Size of Sample Calculated for 1932-33," September, 1932, pages 50-57.

c. "Analysis of Variability in Staple Length of Cotton Ginned During the Seasons of 1931-32 and 1932-33 by Certain Gins and Allocation of Calculated Sample for 1933-34," July 28, 1933, pages 32-33.

d. "Procurement of Samples by the Grade and Staple Section and Suggested Apportionment of the Aggregate of Samples for the 1934-35 Season," April 28, 1934, pages 48-51c.

It is contemplated that more comprehensive reports will be prepared in future years.